EVALUATION OF NATURAL LANGUAGE INTERFACES TO DATA BASE SYSTEMS

Bozena Henisz Thompson
California Institute of Technology

INTRODUCTION

Is evaluation, like beauty, in the eye of the beholder? The answer is far from simple because it depends on who is considered to be the proper beholder. Evaluators may range from casual users to society as a whole, with system builders, sophisticated users, linguists, grant providers, system buyers, and others in between. The members of this panel are system builders and linguists — or rather the two fused into one — but, I believe, interested in all or almost all actual or potential bodies of evaluators. One of our colleagues expressed a forceful opinion while being a member of a similar panel at last year's ACL conference: "Those of us on this panel and other researchers in the field simply don't have the right to determine whether a system is practical. Only the users of such a system can make that determination. Only a user can decide whether the NL [natural language] capability constitutes sufficient added value to be deemed practical Only a user can decide if the system's frequency of inappropriate response is sufficiently low to be deemed practical. Only a user can decide whether the overall NL interaction, taken in toto, offers enough benefits over alternative formal interactions to be deemed practical" [1]. It is hard for me to disagree, since I argued as forcefully on the basis of my study of users' evaluation of machine translation [2] — a study which was prompted by the evaluations of the quality of machine translation as viewed by linguists and users, ranging from 35% acceptable for the former to 90% for the latter. What the study also showed was that the practicality of the output could indeed only be judged by the users, since even incomplete and stylistically very inelegant translations were found quite useful in practice because they, on the one hand, provided, however crudely, the information sought by the users, and, on the other hand, the users themselves brought knowledge that made the texts far more understandable and useful than might appear to a nonspecialist linguist. But this endorsement on my part of the user as the ultimate judge in evaluations does not preclude my fully subscribing to Norm Sondheimer's [3] introductory comments to this panel stating that to "make progress as a field, we need to be able to evaluate." We are now less likely to confuse the issue of the evaluation by people like ourselves and the judgment of the users, less likely to be surprised at the discrepancies, and less likely to be surprised at the users' acceptance of the limitations of our NL interfaces. Also, we are far more aware of the fact that evaluations of "worth" or "quality" have to be conducted in the contexts of the actual, perceived needs. In extensive studies on evaluation of innovations, Mosteller [4], the recently retired president of AAAS, found that "successful innovators better understand user needs; [and] pay more attention to marketing. . . ." The same source, however, leads me to the notorious difficulties of evaluation given the wide range of evaluators and their purposes. We are all undoubtedly convinced of the value of NLI for the society as a whole, but the evaluation of experiments with these interfaces is another matter. Mosteller was faced with social, sociomedical, and medical fields. Let me recount some of the studies he and his team made for reasons which will soon become obvious. His team scored a given program on a scale from plus two to minus two with zero meaning there was essentially no gain. Accordingly, a study of delinquent girls that identified them but failed to prevent them from delinquency received a zero. Likewise, a zero was assigned to a probation experiment for conviction for public drunkenness in which three methods were used: (1) no treatment, (2) an alcoholic clinic, and (3) Alcoholics Anonymous. Since the "no treatment" group performed somewhat better, short-term referrals were considered of no value. A minus one was given to a study whose results were opposite to those hoped for: a major insurance company increased outpatient benefits in the hope of decreasing hospital costs, but the outpatient group's hospital stays increased. Finally, a double plus was awarded to an experiment involving the Salk vaccine, which was, predictably, very successful. Now this kind of evaluation may be justified when the needs of the society are at stake. I have gone into these details, however, for the purpose of expressing the opinion, in which I know I'm not alone, that negative results are as important as positive ones, that evaluation in our case is almost equivalent to the amount of information obtained in an experiment. An experiment whose results would be totally predictable would be almost useless, but one with results different from those hoped for might be embarrassing but very valuable. Another comment prompted by those evaluations is that the application of any rigid, fine scale is totally inappropriate in the case of NLI evaluations.

NLI EVALUATIONS

A. METHODOLOGY AND SOME RESULTS

It had been widely taken for granted some time ago that NLI is as good as is its grammar, and a grammar is as good as it is extensive. The specific needs of users, the requirements of special tasks and the like took a back seat. The nature of human discourse was yet to be explored. Happily, we have been in a different situation for some time. When the REL [5, 6, 7] system was getting into a reasonably sturdy shape with respect to speed and bugs, I started planning experiments to test it. There was important literature about discourse, especially in sociology, such as the work of Schegloff. It was thus clear that successful NLI experiments had to be based on knowledge of human discourse. It was also clear that that was the way to make the interface more natural. This assumption has already been fruitful: the NL interface in POL [9], a successor to REL, has already been extensively improved as a result of the REL-related experiments.

Experiments were made in three modes: in addition to face-to-face and human-to-computer, terminal-to-terminal communication was examined, since at present that is the only practical mode of accessing the computer. Through early 1980, over 80 subjects, 80,000 words, and over 50 hours were analyzed in great detail. In the fall of 1980, another 13 subjects were tested in the computational mode only, adding approximately 20 hours. From the start, the experiments were encouraging, although limited to two modes: F-F and T-T. Interactions not only showed a great deal of structure but extensive similarities in both modes, the most important being the constancy of the number of words in sentences (about 70%); the length of sentences (about 7 words); the existence of fragments (70% of messages in F-F and 50% in T-T containing them); and phatics (10% of total for F-F and 5% for T-T). Thus similarities between the modes were a candidate for consideration in experiments in the computational mode, the T-T mode being seemingly quite far removed from natural F-F. The sentence having historically been the unit of analysis (and since phatics were considered of lesser importance from the computational view, although of great interest in general), my attention turned to fragments. REL allowed for three non-sentence type structures: "NP?" (including number parsed into NP); "all/none or number" answers; and

definitions introducible by the user which make it possible to include individual knowledge and terminology. The analysis of F-F and T-T protocols, however, showed the existence of other fragment categories, finally analyzed into a dozen categories (see [8]). Since they constitute a considerable amount of F-F conversations and even T-T protocols, they clearly had to be watched for in computational experiments.

The experiments for actually observing user-system interaction were conducted in the winter term of 1979/80 and produced 21 protocols, the analysis of which was compared with results of eight F-F and four T-T experiments. Another 13 computational experiments done in the fall confirmed the results of the earlier ones. The task in all three modes was a real one: loading cargo onto a ship, the data coming from the actual environment of loading U.S. navy ships by a group in San Diego, California. In the F-F and T-T experiments, two persons were involved — one given cargo items to be loaded, the other information about decks (details in [8]). In the computational mode (H-C) the ship data was in the computer and the list of cargo to be loaded was handed to the subjects, all with Caltech background. Details being available elsewhere and space limited here, only some major results are given here. Table 1 shows the comparison of the three modes.

TABLE 1

	F-F	T-T	H-C
Sentence length	6.8	6.1	7.8
Message length	9.5	10.3	7.0
Fragment length	2.7	2.8	2.8
% words in sentences	68.8	72.8	89.3
% words in fragments	17.2	21.1	10.7

	Total	Avg.	Total	Avg.	Total	Avg.
Messages	5574	697	310	78	1093	52
Parsed & nonparsed					1615	77
Sentences	5302	663	385	77	882	42
Fragments	3253	402	230	58	211	10
Phatics (including connectors & tags)	4842	605	148	37	46	2

	Total	Total	Total
Words in messages	49800	3285	8525
Words in sentences	34266	2393	6880
Words in fragments	8584	694	823

As can be seen, several statistics show similarities: sentence length, message length, fragment length, percentage of words in sentences and fragments. The closeness of the average of messages in T-T and parsed and nonparsed inputs in H-C is striking.

Table 2 (the meaning of abbreviations is given below the table) deals with fragments. It is mostly self-explanatory, as is the absence of definitions from F-F and T-T (although some abbreviations used there fall in this category) and the absence of some other categories from T-T and H-C. At least two comments, however, are necessary. The surprisingly low use of terse questions in H-C may be accounted for by the tendency toward a formal style in computational interaction. The definitions used were often of quite complex character, although far fewer than could be hoped for due apparently to lack of familiarity with this capability. The complex character of definitions undoubtedly had some effect on the length of sentences in the H-C mode.

TABLE 2

	F-F		T-T		H-C	
	Total	%	Total	%	Total	%
E	532	16.4	10	4.3		
ADD	425	13.1	41	17.8		
CORR	56	1.7				
COMP	95	2.9	2	.9		
SELF	114	3.5				
TR	571	17.6	67	29.1	91	37.8
TQ	411	12.6	31	13.4	67	27.8
TI	297	9.1	48	20.9		
FS	413	12.7	23	10.0	30	12.4
TRUN	339	10.4	9	3.9		
DEF					53	22.0
P	4842		148		46	
C	1936		34			
T	31					

Abbreviations

E (Echo): An exact or partial repetition of usually the other speaker's string. Often an NP, but it may be an elliptical structure of various forms.

ADD (Added Information): An elliptical structure, often NP, used to clarify or complete a previous utterance, often one's own, e.g., "It doesn't say anything here about weight, or breaking things down. Except for crushables.", "It's smaller. 36"x20"x17"." Spelling out words was included here.

CORR (Correction): This may be done by either speaker. If done by the same speaker it is related to false start, but semantic considerations suggest a correction, e.g., "Those are 30, uh, 48 length by 40 width by 14 height."

COMP (Completion): Completion of the other speaker's utterance, distinguished from interruption by the cooperative nature of the utterance, e.g., "A: I've got a lot of...I've got B: two pages. A: Yeah."

SELF (Talking to Oneself): Mutterings, even to the point of undecipherability, not intended for the other person.

TR (Terse Reply): An elliptical reply, often NP, e.g., "No.", "Probably meters.", "50 and 7.62."

TQ (Terse Question): An elliptical question, often NP, e.g., "Why?", "How about pyrotechnics?", "Which ones?"

TI (Terse Information): A rather elusive category, neither question, reply nor command, an elliptical statement but one often requiring an action.

FS (False Start): These are also abandoned utterances, but immediately followed by usually syntactically and semantically related ones, e.g., "They may, they may be identical classes.", "Well, the height, the next largest height I've got is 34."

TRUN (Truncated): An incomplete utterance, voluntarily abandoned.

DEF (Definition): E.g., "Define: ED: each deck of the Alamo."

P (Phatics): The largest subgroup of fragments whose name is borrowed from Malinowski's term "phatic communion" with which he referred to those vocal utterances that serve to establish social relations rather than the direct purpose of communication. This term has been broadened to include all fragments which help keep the channel of communication open, such as "Well", "Wait", and even "You turkey". Two subcategories of phatics are:

C (Dialogue Connectors): Words such as "Then", "And", "Because" (at the beginning of a message or utterance).

T (Tag Questions): E.g., "They're all under 60, aren't they?"

B. SYSTEM PERFORMANCE, SYNTAX USED, SPECIAL STRATEGIES, AND ERROR ANALYSIS

System performance can obviously be evaluated in a number of ways, but without good response time meaningful experiments are impossible. When much data is involved in processing a delay of a few minutes can probably be tolerated, but the vast majority of requests should be responded to within seconds. The latter was the case in my experiments. Fairly complex messages of about 12 words were responded to in about 10 seconds. The system clearly has to be reasonably free of bugs — in my case, 12 bugs were hit in the total of 1615 parsed and nonparsed messages. The adequate extent of natural language syntax is impossible to determine. Table 3 shows the syntax used by my subjects.

TABLE 3

SENTENCE TYPES

	Total	%
All sentences	882	
Simple sentences, e.g., "List the decks of the Alamo."	651	73.8
Sentences with pronouns, e.g., "What is its length?", "What is in its pyrotechnic locker?"	30	3.4
Sentences with quantifier(s), e.g., "List the class of each cargo."	71	8.0
Sentences with conjunctions, e.g. "What is the maximum stow height and bale cube of the pyrotechnic locker of the AL?"	88	10.0
Sentences with quantifier and conjunction(s), e.g., "List hatch width and hatch length of each deck of the Alamo."	23	2.6
Sentences with relative clause, e.g., "List the ships that have water."	6	.7
Sentences with relative clause (or related construction) and comparator, e.g., "List the ships with a beam less than 1000."	6	.7
Sentences with quantifier and relative clause, e.g., "List height of each content whose class is class IV."	2	.23
Sentences with quantifier, conjunction and relative clause, e.g., "List length, width and height of each content whose class is ammunition."	2	.23
Sentences with quantifiers and comparator, e.g., "How many ships have a beam greater than 1000?"	3	.34
Wh-questions		75.0
Yes/no questions		1.0
Commands		19.0
Statements (data addition)		5.0

Considering the wide range of REL syntax [7], the paucity of complex sentences is surprising. The use of definitions which often involved complex constructions (relative clauses, conjunctions, even quantifiers) had a definite influence. So did, undoubtedly, the task situation causing optimization of work methods. The influence of the specific nature of the task would require additional studies, but the special device provided by the system (a loading prompt sequence — which was not analyzed) was employed by every subject. Devices such as these obviously are a great aid in accomplishing tasks. They should be tested extensively to determine how they can augment the naturalness of NLIs. Other reasons for the relatively simple syntax used were special strategies: paraphrasing into simpler syntax even though a sentence did not parse for other reasons; "success strategy" resulting in repetitious simple sentences; or possibly just "baby talk" due to the suspicion of the computer's limitations.

An interesting fact to note is that similar results with respect to syntax were obtained in the experiments with USL, the "sister system" of REL developed by IBM Heidelberg [10] — with German used as NLI in two studies of high school students: predominance of wh-questions (317 in total of 451); not many relative clauses (66); commands (35); conjunctions (26); quantifiers (15); definitions (11); comparisons (2); yes/no questions (1).

An evaluation which would not include an analysis of unparsed input would at best be of limited value. It was shown in Table 1 that 1093 out of 1615 or about two thirds were parsed in my experiments.

TABLE 4

	Total	%
Vocabulary	161	36.1
Punctuation	72	16.1
Syntax	62	13.9
Spelling	61	13.6
Transmission	32	7.2
Definition format	30	6.7
Lack of response	16	3.6
Bug	12	2.7

Table 4 summarizes the categories of errors. The predominance of vocabulary is not surprising, but relatively few syntactic errors are. In part this may be due to the method of scoring in which errors were counted only once, so if a sentence contained an unknown vocabulary item (e.g. "On what decks of the Alamo may cargo be stored?") but would have failed on syntactic grounds as well, it would fall in the vocabulary category. A comparison can be made here with Damerau's study [11] of the use of the TQA system by the city planning department in White Plains, at least with regard to the total of queries to those completed: 788 to 513. So, again, roughly two thirds were parsed. In other categories "parsing failure" is 147, "lookup failures" 119, "nothing in data base" 61, "program error" 39, but this only points to the general difficulties of comparisons of system performance.

SOME CONCLUSIONS

Norm Sondheimer suggested some questions we might try to answer. What has been learned about user needs? What most important linguistic phenomena to allow for? What other kinds of interactions? Error analysis points in the obvious directions of user needs, and so do the types of sentences employed. While it is justified to quit the search for an almost perfect grammar, it would be a mistake to constrain it to the constructions used. Improved naturalness can be achieved with diagnostics, definitions, and devices geared to specific tasks such as special prompting sequences. Some tasks clearly require math in the NLI. How good are systems? An objective measurement is probably impossible, but the percentage of requests processed might give some idea. In the case of a task situation such as loading cargo items, the percentage of task completion may signal both system performance and user satisfaction. System response times are a very important measure. The questionnaire method can and has been used (in the case of MT and USL), but as yet there is too little experience to measure user satisfaction. Users seem very good at adapting to systems. They paraphrase, use success strategy, simplify syntax, use special devices — what they really do is maximize their performance with respect to a given task.

What have we learned about running evaluations? It is important to know what to look for, therefore the need for good knowledge of human to human discourse. Good system response times are a sine qua non. Controlled experiments have the advantage of being replicable, a crucial factor in arriving at evaluation criteria. Determining user bias and experience may be important, but even more so is user training. Controlled experiments can show what methods are most effective (e.g. a manual or study of protocols?). Study of user comments — phatic material — gives some measure of user (dis)satisfaction (I have seen "You lie," but I have yet to see "Good boy, you!"). Clearly, the best indication of user satisfaction is whether he or she uses the system again. Extensive long-term studies are needed for that.

What should the future look like? Task oriented situations seem to be a promising environment for NLI. The standards of NL systems performance will be set by the users. Future evaluations? As Antoine de Saint-Exupéry wrote, "As for the Future, your task is not to foresee, but to enable it."

REFERENCES

1. Harris, Larry R. "Prospects of Practical Natural Language Systems." Proceedings of the 18th Annual Meeting of the Association for Computational Linguistics, June 1980, p. 129.

2. Henisz-Dostert, B.; Macdonald, R. R.; and Zarechnak, M. Machine Translation. The Hague: Mouton, 1979.

3. Sondheimer, N. K. "Evaluation of Natural Language Interfaces to Data Base Systems." Proceedings of the 19th Annual Meeting of the Association for Computational Linguistics, June 1981.

4. Mosteller, F. "Innovation and Evaluation." Science (February 27, 1981):881-886.

5. Thompson, F. B. and Thompson, Bozena H. "Practical Natural Language Processing: The REL System as Prototype." In Advances in Computers, ed. M. Rubinoff and M. C. Yovits. Vol. 13. New York: Academic Press, 1975.

6. Thompson, Bozena H. and Thompson, F. B. "Rapidly Extendable Natural Language." Proceedings of the 1978 National Conference of the ACM, pp. 173-182.

7. Thompson, Bozena H. REL English for the User. Pasadena: California Institute of Technology, 1978.

8. Thompson, Bozena H. "Linguistic Analysis of Natural Language Communication with Computers." COLING 80: Proceedings of the 8th International Conference on Computational Linguistics, Tokyo, October 1980, pp. 190-201.

9. Thompson, Bozena H. and Thompson, F. B. "Shifting to a Higher Gear in a Natural Language System." Proceedings of the National Computer Conference, May 1981.

10. Lehmann, Hubert; Ott, Nikolaus; Zoeppritz, Magdalena. "User Experiments with Natural Language for Data Base Access." COLING 78: Proceedings of the 7th International Conference on Computational Linguistics, Bergen, August 1978.

11. Damerau, Fred J. The Transformational Question Answering (TQA) System: Operational Statistics - 1978. RC 7739. Yorktown Heights: IBM T. J. Watson Research Center, June 1979.